HiScene: Creating Hierarchical 3D Scenes with Isometric View Generation

Wenqi Dong[1,2‡] Bangbang Yang[2*] Zesong Yang[1,2] Yuan Li[1]
Tao Hu[2] Hujun Bao[1] Yuewen Ma[2] Zhaopeng Cui[1†]
[1]Zhejiang University [2]ByteDance

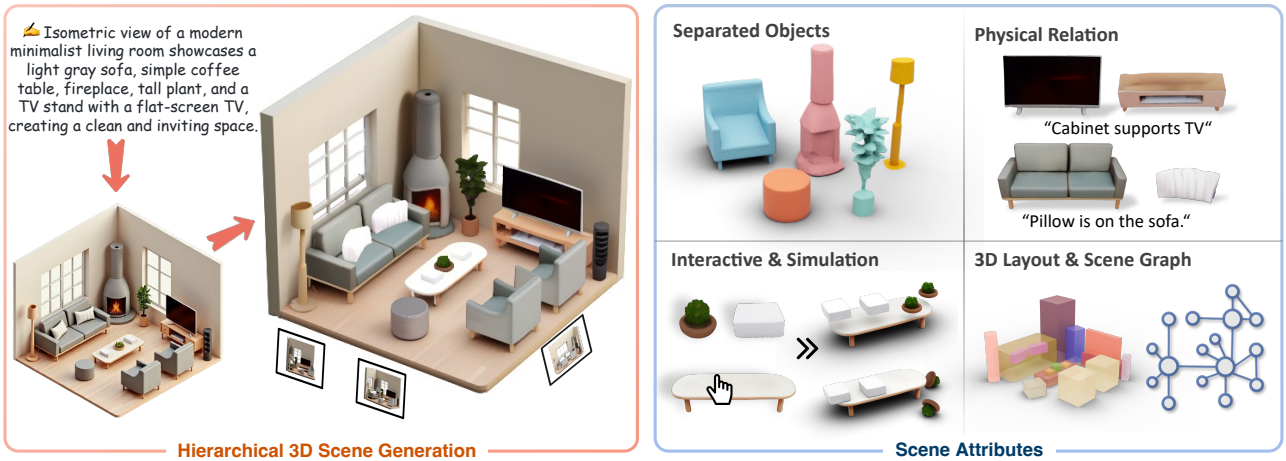


Figure 1. **HiScene** allows users to generate scene-level 3D assets with natural layout and appealing looking, while delivering compositional items for versatile applications such as interactive editing and simulation.

Abstract

Scene-level 3D generation represents a critical frontier in multimedia and computer graphics, yet existing approaches either suffer from limited object categories or lack editing flexibility for interactive applications. In this paper, we present HiScene, a novel hierarchical framework that bridges the gap between 2D image generation and 3D object generation and delivers high-fidelity scenes with compositional identities and aesthetic scene content. Our key insight is treating scenes as hierarchical "objects" under isometric views, where a room functions as a complex object that can be further decomposed into manipulatable items. This hierarchical approach enables us to generate 3D content that aligns with 2D representations while maintaining compositional structure. To ensure completeness and spatial alignment of each decomposed instance, we develop a video-diffusion-based amodal completion technique that effectively handles occlusions and shadows between objects, and introduce shape prior injection to ensure spatial coherence within the scene. Experimental results demonstrate that our method produces more natural object arrangements and complete object instances suitable for interactive applications, while maintaining physical plausibility and alignment with user inputs. More details at https://zju3dv.github.io/hiscene/.

1. Introduction

Recently, we have witnessed remarkable breakthroughs of generative techniques in both 2D and 3D content creation, which empowers users to create stunning images and intricate 3D objects with simple text prompts. However, extending such generation capabilities to scene-level 3D content—for instance, generating complete 3D room models with coherent geometry and appearance guided by user-provided text or images—remains a significant challenge. Recent approaches typically rely on large language models (LLMs) with handcrafted rules to generate 3D scene layouts and place objects accordingly, yet these results often lack realism (e.g., exhibiting constrained object diversity and simplistic arrangements) due to LLMs' limited spatial understanding capabilities [66]. Others attempt to lift 2D images

*Authors contributed equally.
†Corresponding author.
‡Work done during an internship at PICO, ByteDance.

into 3D scenes using depth-based mesh deformation [13, 16]. While these approaches offer category-free generation capabilities, they typically produce scenes as inseparable wholes, limiting interactive applications such as scene editing, object manipulation, and data curation for robotic semantic understanding.

Based on these observations, we believe an ideal scene generation method should have the following properties: **1) Realistic layout & assets**: it should generate scenes with natural object arrangements and diverse content beyond just a few simple objects from limited categories. The scenes should reflect real-world spatial relationships and object interactions. **2) Compositional & complete instances**: for interactive applications, each object in the scene should be a complete, intact 3D entity that can be individually manipulated, edited, or replaced without disrupting the entire scene. **3) Spatial alignment & plausibility**: the generated scene should faithfully represent the user's text or image prompt while maintaining physical coherence and plausibility in the 3D space (e.g., without distorted shapes, unrealistic proportions, or floating placement).

In this paper, we propose a novel hierarchical scene generation framework, named HiScene. Rather than determining how scenes are built in 3D space with handcrafted rules, we leverage the complementary knowledge embedded in image generation models about how scenes should appear with aesthetic appeal and reasonable layout, and instantiate the concrete 3D representation that aligns with the image in a top-down manner. Our key insight is that we can treat a scene as hierarchical level "objects" under the **isometric view**. From the generator's perspective, a room can be seen as a complex object itself, while each individual item within the room can also be separately generated and manipulated. By leveraging this hierarchical approach, HiScene bridges the gap between object-level and scene-level generation, producing complete scenes that benefit from pre-trained object-centric generation priors while maintaining compositional structure. While the hierarchical scene generation approach is technically plausible, we identify several challenges in creating high-fidelity and compositional 3D scenes that we address in this paper.

Hierarchical Scene Parsing. To bootstrap the hierarchical scene generation, we first initialize the entire scene with a pre-trained 3D generation model [70] from the given isometric view image. Once obtaining the complete 3D Gaussian representation [23] of the scene, a key challenge is to accurately isolate individual objects from the scene structure. To address this, we implement a hierarchical scene parsing approach based on "analysis by synthesis" to identify distinct objects and prepare them for subsequent generative refinement. Specifically, we render multi-view images and perform 3D identity segmentation using 2D segmentation priors enhanced with contrastive learning techniques [78]. Then, for each identified object, we render object-centric circular views and carefully identify occlusion regions, enabling us to understand the spatial relationships between objects and more effectively reconstruct complete object identities in the following steps.

Instance Refinement with Video-diffusion-based Amodal Completion. A key challenge during identity refinement is that the rendered instance views often exhibit significant occlusions. Despite advances in 3D object generation, reconstructing complete objects from occluded views remains ill-posed, while directly applying standard inpainting methods often produces implausible results due to limited object understanding (see Fig. 7). Moreover, the target instance might also include ambient shadow caused by the foreground occluder, which cannot be addressed with conventional inpainting frameworks. To tackle this problem, we reformulate the instance refinement as a 2D amodal completion and 3D regeneration task, and propose a novel video-diffusion-based completion framework to handle it. Specifically, our approach treats the amodal completion process as a temporal transition video effect, where occlusions gradually dissolve to reveal the complete object. To enable this capability, we construct a specialized dataset for training video diffusion model to perform such completion transitions. The temporal nature of our video-based completion effectively handles challenging cases including occlusion shadow removal, outperforming static image-based inpainting or completion methods by preserving structural coherence and producing more plausible results even in complex scenarios.

Spatial Aligned Generation. Even with advanced 3D generation models to refine each segmented identity, ensuring spatial alignment between the refined object and its original placement is non-trivial due to the unposed nature of compressed latent [70]. As a result, naïvely applying refinement with amodally completed object views might produce objects with variant shapes, making them incompatible with the original scene layout. To address this issue, we propose a shape prior injection mechanism that conditions the refinement stage of each identity. Specifically, we first extract a geometric shape prior from a view-aligned generation method [72], and use this aligned shape prior as the latent initialization for our refinement pipeline rather than starting from random noise. This approach significantly reduces geometric ambiguity during refinement and ensures proper spatial alignment between the generated objects and the original scene context.

The contribution of the paper can be summarized as follows. **1)** We propose HiScene, a novel scene-level generation method that produces high-fidelity 3D scenes with compositional identities, natural scene arrangement and diverse content. HiScene bridges the gap between 2D and 3D generation by leveraging isometric views, enabling effective hierarchical scene-level generation with pre-trained 3D object generation models. **2)** We develop an analysis-by-synthesis

approach for scene parsing with zero-shot 3D semantic segmentation, and introduce a video-diffusion-based amodal completion method that effectively handles occlusions and ambient shadows during instance refinement. **3)** We design a spatial alignment mechanism with shape prior injection that ensures refined objects maintain proper geometric alignment with the original scene context, ensuring the object coherence and physical plausibility of the compositional 3D scene. **4)** Extensive experiments demonstrate the effectiveness of our approach, demonstrating the superior performance of video-diffusion-based amodal completion in handling complex occlusions and shadows, and the high-quality scene decomposition and object refinement across various challenging scenarios.

2. Related Work

2.1. 3D Object Generation

Motivated by the recent advances of diffusion techniques in 2D image generation, numerous works [9, 34, 36, 46, 53, 61, 65, 67, 77] are being made to apply diffusion models to 3D object generation. DreamFusion [53] first attempts to distill 2D gradient priors from the denoising process by employing score distillation sampling loss (SDS loss). Follow-up methods aim to enhance both the quality [9, 36, 46, 65, 67] and efficiency [34, 61, 77] of this method. Zero123 [41] constructs paired viewpoint data on the large-scale dataset Objaverse [15] and fine-tunes the 2D latent diffusion model to achieve arbitrary novel view synthesis and 3D generation under single-image condition. Later works [42, 43, 59, 60, 64] address the issue of cross view inconsistency in Zero123 by synchronously generating multiview images. To enhance 3D object generation efficiency, Large Reconstruction Model (LRM) [17] employ a transformer-based architecture that directly reconstructs 3D objects from single-view image through feed-forward inference. Some methods [31, 40, 62, 71, 72, 74] incorporated multi-view information as input, significantly improving generation quality. Recent native 3D generation model [11, 28, 32, 70, 85, 87, 90, 91] such as Shape2Vecset [85], CLAY [87], and TRELLIS [70] adopt a decoupled strategy, dividing the process into geometric structure generation and texture generation stages. These methods train generative model directly on 3D data rather than traditional multi-view data, substantially improving the geometric accuracy and consistency of generated results. Our method use existing native 3D generation model, treating scenes as hierarchical level of "objects" to ultimately achieve interactive scene generation.

2.2. Text-to-3D Scene Generation

Text-conditioned 3D scene generation has advanced rapidly in recent years. Text2Room [16] and follow-up works [13, 79, 80, 86, 92] leverage latent diffusion models [58] and monocular depth estimators [4, 6] to iteratively generate textured 3D meshes or Gaussian Splatting scenes by inpainting from randomly sampled camera angles. However, these methods typically produce coupled scenes where object instances are difficult to isolate. Some methods [35, 48, 50, 63, 75, 81, 82] first generate 3D scene layouts, then obtain individual objects through retrieval or generation methods, placing them within the scene according to the generated layout. BlockFusion [69] and its follow-up methods [8, 45] adopt a native 3D scene generation strategy, dividing the scene into multiple blocks and expanding these blocks into a complete scene during generation. However, these two types of methods are often limited to specific categories from training data. Recent methods [29, 33, 88, 93] explore interactive general scene generation using Large Language Models (LLMs) to construct 3D layouts, while generating individual objects through Score Distillation Sampling-based optimization. However, LLMs still lack sufficient spatial understanding capabilities, making it difficult to generate complex and physically plausible 3D layout structures. By contrast, our method adopts a top-down hierarchical generation that ensures global layout and appearance coherence while maintaining the separability of individual objects.

2.3. Image Inpainting and Amodal Completion

Image Inpainting and Amodal Completion are classic problems in computer vision. Image inpainting aims to restore masked regions in images with reasonable and natural content, requiring specification of the inpainting area. Traditional methods [5, 14, 39, 51, 89] offen relied on auxiliary hand-engineered features with often poor results. Recent approaches [2, 3, 10, 19, 20, 38] utilize diffusion models for text-guided completion, typically regenerating content in masked regions while preserving the rest of the image. In contrast, Amodal Completion aims to generate the complete form of an object from its visible view. Traditional methods [18, 21, 22, 37, 54, 57, 83] focused on generating complete segmentation masks or predicting bounding boxes, while recent zero-shot diffusion-based method [49, 73] further recover the complete content. We adopt Amodal Completion rather than Image Inpainting because Image Inpainting requires manually specified inpainting regions, whereas in Amodal Completion, the observed regions can be automatically obtained through existing segmentation networks, which is crucial for our automated generation of candidate images for each object during the hierarchical scene parsing stage.

3. Method

We present HiScene, a novel hierarchical framework for generating compositional 3D scenes with intact and manipulatable objects. As illustrated in Fig. 2, we first initialize a 3D Gaussian Splatting scene from a generated isometric view.

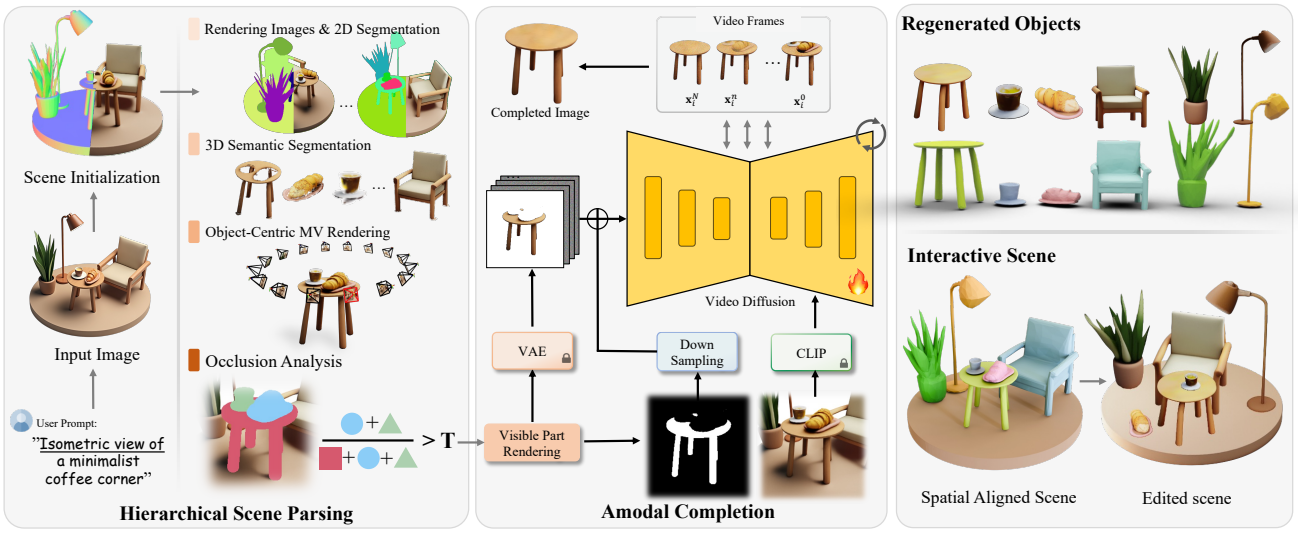


Figure 2. **Overview of HiScene.** Our hierarchical framework generates 3D scenes with compositional identities through three main stages. First, we create a 3D scene from a generated isometric view. Next, we perform scene parsing to obtain precise object segmentation, followed by multi-view rendering and detailed occlusion analysis for each identified instance. Finally, we apply our video-diffusion-based amodal completion to generate complete views of each instance, which serve as guidance for regenerating intact objects with proper spatial alignment in the scene. The resulting 3D scene features fully compositional identities, facilitating user-directed modifications like interactive scene editing.

(a) Perspective view (b) Isometric view

Figure 3. Comparison of perspective view and isometric view of a living room scene. Zoom in for more details.

We then perform hierarchical scene parsing with semantic segmentation to identify distinct objects and obtain each object' multi-view rendering and occlusion analysis. Finally, we conduct video-diffusion-based amodal completion to address object occlusion and generate intact, spatially-aligned objects that enable interactive scene manipulation.

3.1. Preliminary

Isometric View. In computer graphics, isometric view is an orthographic projection method used to render 3D scenes into images. Isometric view offers three key advantages for scene-level generation: 1) *Distortion-free*: Unlike perspective projection, isometric view maintains consistent proportions without perspective distortion, ensuring accurate object representation. 2) *Minimal-occlusion*: As shown in Figure 3, isometric view captures scenes from an elevated angle, revealing multiple faces of objects simultaneously while minimizing occlusion between scene elements. 3) *Scene-as-object*: The unified representation of scenes in isometric view allows the entire scene to be treated as a cohesive entity, enabling direct generation with object-centric generative models.

Native 3D Generation Model. We leverage native 3D generation models TRELLIS [70] to bootstrap hierarchical scene generation. TRELLIS introduces a unified Structured LATent (SLAT) representation $\boldsymbol{z}$ to characterize a 3D asset $\mathcal{O}$, as:

$$\boldsymbol{z} = \{(\boldsymbol{z}_i, \boldsymbol{p}_i)\}_{i=1}^{L}, \quad \boldsymbol{z}_i \in \mathbb{R}^C, \ \boldsymbol{p}_i \in \{0, 1, \ldots, N-1\}^3, \quad (1)$$

TRELLIS employs a two-stage generation pipeline. In the first stage, it generates the sparse structure $\{\boldsymbol{p}_i\}_{i=1}^{L}$ of $\mathcal{O}$ by first using a transformer model $\boldsymbol{\mathcal{G}}_{\mathrm{S}}$ to produce a low-resolution feature grid $\boldsymbol{S} \in \mathbb{R}^{D\times D\times D\times C_{\mathrm{S}}}$, followed by a latent feature decoder $\mathcal{D}_S$ to obtain a dense binary 3D grid $\boldsymbol{O} \in \{0,1\}^{N\times N\times N}$. The grid $\boldsymbol{O}$ is then converted to the set of 3D coordinates $\{\boldsymbol{p}_i\}_{i=1}^{L}$. In the second stage, TRELLIS uses another transformer model $\boldsymbol{\mathcal{G}}_{\mathrm{L}}$ to generate the corresponding structure features $\{\boldsymbol{z}_i\}_{i=1}^{L}$ for these coordinates. The complete SLAT representation $\boldsymbol{z}$ is then processed through specialized decoders ($\mathcal{D}_{\mathrm{NeRF}}$, $\mathcal{D}_{\mathrm{Mesh}}$, or $\mathcal{D}_{\mathrm{GS}}$) to produce the final 3D asset $\mathcal{O}$ in various formats (NeRF, meshes, or 3DGS).

3.2. Hierarchical scene parsing

We define an interactive scene $\mathcal{S} = \{\{\mathcal{O}_i, \mathcal{C}_i\}_{i=1}^{N}\}$ containing multiple separable complete objects $\{\mathcal{O}_i\}_{i=1}^{N}$ represented by 3DGS, with their configurations $\mathcal{C}_i = \{p_i, r_i, s_i, l_i\}$, where each configuration includes position $p_i \in \mathbb{R}^3$, rotation $r_i \in SO(3)$, scaling $s_i \in \mathbb{R}^3$, and semantic label l_i. To obtain this, as shown in Figure 2, we first perform hierarchical scene parsing.

Scene Initialization Given users' text prompt T, we first generate isometric view candidate images by prepending a fixed prefix *"Isometric view of "* to the T. Then, for the selected scene image I_{scene}, we obtain the initial scene representation $\mathcal{S}_0$ through TRELLIS. We adopt 3D Gaussian Splatting as our scene representation method.

3D Semantic Segmentation. We employ OmniSeg3D-GS [78], a contrastive learning-based semantic segmentation method for 3DGS. Specifically, after obtaining the initial representation $\mathcal{S}_0$, we render multi-view images $\mathcal{I} = \{I_j\}_{j=1}^{N_{VS}}$ from predefined viewpoints $\{V_j\}_{j=1}^{N_{VS}}$, where N_{VS} is the number of scene views. Unlike OmniSeg3D-GS, which aims to achieve omniversal segmentation, our goal is to obtain separated objects $\{O_i\}_{i=1}^{N}$. Therefore, instead of using SAM [25], we employ EntitySeg [55], an off-the-shelf entity segmentation network, to generate class-agnostic instance-level 2D segmentation masks $\mathcal{M} = \{M_j\}_{j=1}^{M}$ for the multi-view images.

Object-Centric Multiview Rendering. As demonstrated in Figure 2, due to the occlusion between objects in 3D scenes, after 3D segmentation, some objects (such as tables, croissants, coffee cups, etc.) typically appear incomplete. To further recover complete objects, we apply the following method to each object $\mathcal{O}_i$ in the scene: First, in the object's local coordinate system, we render multiview images $\{I_j^i\}_{j=1}^{N_{VO}}$ from predefined viewpoints $\{V_j^i\}_{j=1}^{N_{VO}}$, where N_{VO} is the number of object views. These rendered images will be served as candidate images for further object-centric instance regeneration.

Occlusion Analysis. After obtaining candidate images $\{I_j^i\}_{j=1}^{N_{VO}}$ for each object $\mathcal{O}_i$, we need to evaluate whether $\mathcal{O}_i$ is occluded by other objects in these images. For each candidate image I_j^i, we employ an advanced Vision Language Model (VLM) for occlusion detection. If occlusion is detected, we calculate the occlusion ratio of the target object. When this ratio exceeds threshold τ, we apply Amodal Completion to recover the visual information in the occluded regions, thereby enhancing the quality and completeness of the candidate images.

For more details, please refer to the supplementary materials.

3.3. Amodal Completion

Task Definition. As discussed in the Section 2.3, unlike image inpainting, amodal completion task aims to recover the complete and plausible form of the object when provided with an input image I and visible mask M (i.e., objects' visible part as shown in Figure 4 (c)). Unlike existing method Pix2gestalt[49] that fine-tune image generation models, we define this task as a temporal transition video effect, where occlusions gradually dissolve to reveal the complete object. Video models, trained on large-scale high-quality data, can learn temporal changes in the real world, thus possessing stronger prior knowledge that helps more accurately infer the complete form of occluded parts.

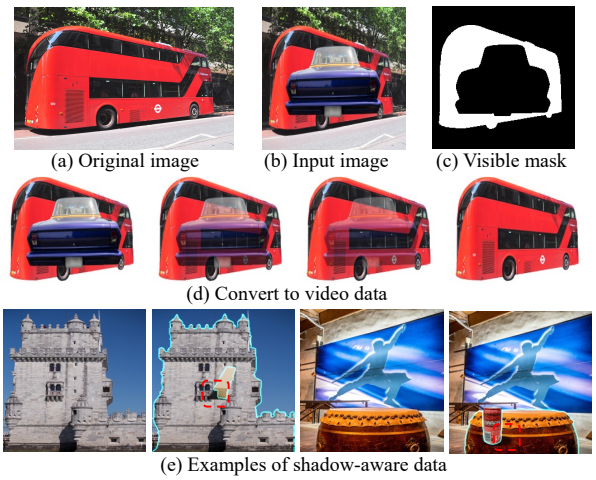


Figure 4. We present an data curation example of amodal completion, including original image (a), occluded input image (b), visible mask (c), and the linear blended video (d). We also present shadow-aware data examples (e).

Dataset curation. During object completion, apart from filling occluded parts, we need to remove notable visual artifacts caused by occlusion, such as shadows. Objects directly segmented from the SA-1B [24] dataset cannot meet the requirements for constructing data with shadow effect. To this end, we constructed a large-scale dataset of objects with shadow occlusions using synthetic data. By filtering the Objaverse dataset [15], we obtained 181K high-quality 3D objects. For each object, we utilized rigid body simulation to naturally place objects on the ground, ensuring realistic shadow effects. Additionally, we configured random lighting setups and employed the path-tracing renderer to generate 468K synthetic images. Integrating the data from Pix2gestalt, we ultimately built a training set comprising 1.32 million image pairs. Then, we convert image pair into the required video data. As shown in Figure 4 (d), we adopt a linear blending approach to put foreground object over the complete objects in image planes.

As shown in Figure 2, we employ the Image-to-Video model Stable Video Diffusion [7]. Specifically, the visible part I_{vis} is encoded through a VAE and used as the first

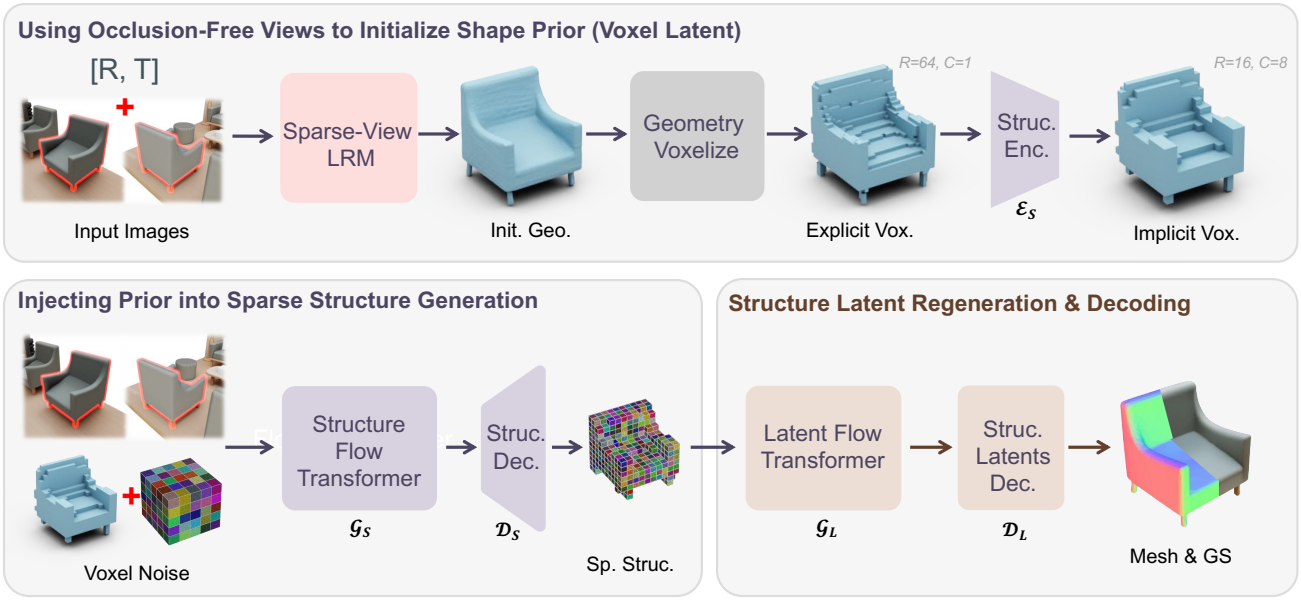


Figure 5. An illustration of Spatial Aligned Generation. We use sparse-view LRM to initialize spatial aligned shape prior (voxel latent), and inject this prior by initializing voxel noises upon it during native 3D generation, thus ensuring regenerated assets adhering the original scene.

frame input, concatenated with the downsampled visible region mask M along the feature dimension. The whole image I is processed through CLIP to extract features, which are then injected into the model via cross-attention. During training, we add noise $\epsilon \sim \mathcal{N}(0, I)$ to the original data x_0 to obtain $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t}\epsilon$, and then use the network to predict the noise $\epsilon_\theta(x_t, t)$, minimizing the following loss function:

$$\mathcal{L}_{dm} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t, I, I_{vis}, M)\|^2\right], \quad (2)$$

Once the Amodal completion model is trained, we apply it to object candidate images with occlusion, which recovers intact image inputs for Spatial Aligned Generation stage.

3.4. Spatial Aligned Generation

After obtaining objects' occlusion-free views, we aim to regenerate each object to achieve intact instances while preserving their original scale and poses. However, directly applying regeneration using native 3D generative models [70] F_{Native} often results in canonical objects that lose alignment with the original scene context. To address this limitation, we propose injecting spatially aligned shape priors derived from multi-view large reconstruction models F_{LRM} [72] (refer to it as LRM for clarity) into the native 3D generation process.

As shown in Figure 5, for each incomplete objects $\mathcal{O}_i$, we first use F_{LRM} to reconstruct an initial geometric structure S_{init} from the input observation views and their corresponding camera parameters $[r_i, p_i]$, and obtain an explicit voxel representation V through voxelization. Subsequently, we employ the encoder $\mathcal{E}_S$ from the TRELLIS structure generation stage to compress V into a low-dimensional latent feature S_{implicit}. This coarse 3D structure representation serves as guidance for the subsequent refinement process. Specifically, we use S_{implicit} as the voxel latent initialization and add noise corresponding to an intermediate timestep t of the rectified flow model (typically choosing $t \in [0.2, 0.4]$), rather than starting generation from pure random noise. The generator $\mathcal{G}_{\text{S}}$ then produces an optimized structure $\widehat{S}$. Under the guidance of S_{implicit}, the finally generated $\widehat{S}$ not only preserves the configuration of the initial geometry M but also significantly enhances geometric details and texture quality.

4. Experiments

In this section, we first demonstrate method's capabilities of text to interactive 3D scene generation in Sec. 4.1. Next, we evaluate the effectiveness of our method in amodal completion task in Sec. 4.2. Finally, we conduct ablation studies to analyze the different components within our framework in Sec. 4.3.

4.1. Interactive Scene Generation

We compare our method with two state-of-the-art decoupable scene generation methods: GALA3D [93] and DreamScene [29]. GALA3D employs large language models for generating initial layouts, integrates layout-guided Gaussian representation and adaptive geometry control, and utilizes a compositional optimization mechanism. DreamScene introduces Formation Pattern Sampling (FPS) to balance semantic information and shape consistency, and a three-stage camera

Table 1. We perform quantitative evaluation and user studies on the Interactive Scene Generation task.

Method	Ours	GALA3D [93]	DreamScene [29]
↑ Aesthe. Score [47]	**5.46**	4.74	4.71
↑ ImageReward [68]	**-0.28**	-1.67	-0.73
↑ CLIP Score$_{\%}$ [56]	**26.07**	23.50	21.91
↑ Matching Degree	**2.90**	1.76	1.40
↑ Overall Quality	**2.76**	1.75	1.73

sampling strategy to improve the quality of scene generation. However, both methods require predefined 3D layouts as input, which presents a significant barrier for novice users who may find creating reasonable layouts challenging. Large language models often make errors in layout generation as well. As shown in Figure 6, our method addresses these limitations by providing a more intuitive and user-friendly approach to 3D scene generation without requiring explicit layout specifications.

Qualitative Comparison In Figure 6, we present both complete generated scenes and individual objects. As shown, scenes and objects generated by GALA3D and DreamScene exhibit artifacts. The layouts produced by these methods often violate physical constraints and common sense spatial relationships. Additionally, individual objects frequently suffer from oversaturation and the Janus problem. In contrast, our method generates complex yet plausible scenes with individual objects of significantly higher quality compared to the other approaches.

Quantitative Analysis To quantitatively evaluate our method, we employ CLIP Score to assess text-scene alignment, and use ImageReward and Aesthetic Score to evaluate the overall generation quality. As shown in Table 1, our method achieves the best overall performance. DreamScene's more severe multi-face object issues negatively impact its overall scores. Our approach demonstrates superior performance across all metrics, confirming the effectiveness of our layout-free scene generation paradigm.

User Study We also conducted a user study to compare our method against existing approaches. The evaluation focused on two aspects: text-scene alignment and overall quality. We collected 12 different scenes and asked 20 users to rate them on a scale from 1 to 3, with higher scores indicating better results. As shown in Table1, our method achieved the highest ratings, confirming the superior performance of our approach from a human perception perspective.

4.2. Amodal Completion

We assess the performance of amodal segmentation with existing zero-shot methods. Following [49, 84], we evaluate segmentations on Amodal COCO (COCO-A) [94] and Amodal Berkeley Segmentation (BSDS-A) [44] datasets using mean intersection-over-union (mIoU). The COCO-A dataset offers 13,000 amodal annotations across 2,500 images, while the BSDS-A dataset includes 650 objects from 200 images. For both datasets, we evaluate methods that take an image and a (modal) mask of the visible portion of an object as input, and produce an amodal mask representing the full extent of the object. We use the same method as in pix2gestalt [49] to convert the amodal completions into semantic masks.

Table 2. Comparisons with zero-shot methods.

Zero-shot Method	COCO-A	BSDS-A
SAM [25]	60.27	60.20
SD-XL Inpainting [52]	70.08	66.57
Pix2gestalt [49]	82.59	79.59
Ours	**83.84**	**79.80**

We compared our approach with the state-of-the-art method pix2gestalt and two other zero-shot methods, as shown in Table 2. Our method achieves state-of-the-art performance on both datasets, demonstrating that our video model-based approach more effectively recovers occluded objects, resulting in superior segmentation outcomes. We also conducted qualitative experiments on everyday scenes, as illustrated in Figure 7. Our method successfully recovers occluded objects and effectively removes shadows caused by occlusions. While pix2gestalt can reconstruct reasonable shapes, it often produces darkened textures in shadowed regions, likely due to the absence of shadow considerations in its training data. SD-XL Inpainting tends to be influenced by mask boundaries, frequently generating completions that conform to the mask but are semantically unreasonable.

4.3. Ablation Studies

Image vs. video model in amodal completion. To qualitatively compare the capabilities of image and video models, we trained both types of models using data constructed by Pix2gestalt on the SA-1B dataset [25]. We evaluated the overall quality of the generated completions using Aesthetic Score, Q-Align IAA, and IQA metrics, and measured text-image alignment using CLIP Score. As shown in Table 3, under the same data settings, the video model outperforms the image model across all metrics. We attribute this superior performance to the video model's powerful prior knowledge of object continuity and temporal consistency, which enables it to better understand and complete occluded objects with more coherent and realistic results.

Shadow-aware completion for object generation. As illustrated in Figure 8, our observation of the chair is incomplete due to occlusions. To demonstrate the importance of proper amodal completion, we first conducted an ablation experiment where we attempted object generation without amodal

Figure 6. We compare the Interactive Scene 3D generation with GALA3D and DreamScene.

completion. Since object generation models are trained on complete observations, when presented with partial inputs, the model generates incorrect geometric structures based on the incomplete mask contours, often resulting in black textures in the missing regions. Similarly, when amodal completion is performed but shadow artifacts remain, the generated results still exhibit the aforementioned black geometric errors. Our shadow-aware amodal completion method effectively addresses these issues by properly handling both occlusions and shadows, resulting in geometrically accurate and visually coherent object reconstructions.

Spatial Alignment. We finally evaluate the effectiveness of our spatial aligned generation by comparing it with two alternatives: direct native 3D generation (F_{Native} only) and standalone LRM generation (F_{LRM} only). As shown in Figure 9, without spatial alignment, native 3D generation produces objects with incorrect orientation and positioning relative to the ground-truth instance, while LRM generation alone results in compromised appearance fidelity. By leveraging LRM's spatial alignment capabilities as a shape prior for native 3D generation, our approach achieves precise scale

Table 3. We evaluated the effectiveness of the video model.

Datasets & Method	COCO-A		BSDS-A	
	I2I	I2V	I2I	I2V
↑ Aesthe. Score [47]	4.16	**4.30**	4.17	**4.38**
↑ Q-Align IAA% [68]	12.65	**23.48**	14.59	**23.46**
↑ Q-Align IQA% [68]	35.43	**45.80**	35.06	**44.42**
↑ CLIP Score% [56]	20.27	**20.78**	20.63	**21.20**

Figure 7. **In-the-wild Amodal Completion and Segmentation.**

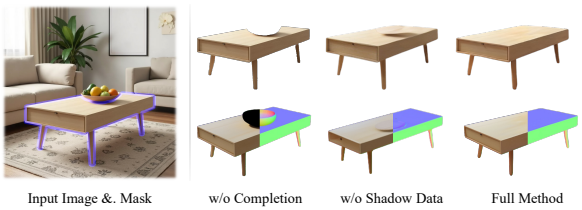


Figure 8. We analyzed the necessity of shadow-aware amodal completion.

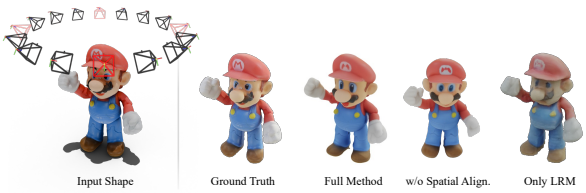


Figure 9. We analyze the effectiveness of Spatial Aligned Generation.

and pose matching with the original scene while preserving rich appearance details and visual quality.

5. Conclusion

In this paper, we have presented HiScene, a novel hierarchical framework for generating compositional 3D scenes. By treating scenes as hierarchical compositions of objects under isometric views, we enable effective scene-level synthesis using pretrained object generation models. To ensure completeness of each object identities, we use video-diffusion-based amodal completion and spatial alignment to aid the regeneration, ensuring spatial coherence within the scene.

Limitations and future works. Despite our advances, scenes generated by HiScene have textures with baked lighting, lacking PBR materials for modern rendering pipelines. Future work will focus on training the generative model to generate scenes that support PBR textures.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 14

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 3

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42(4): 1–11, 2023. 3

[4] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. *arXiv preprint arXiv:2210.03676*, 2022. 3

[5] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 3

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 3

[7] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 5, 16

[8] Alexey Bokhovkin, Quan Meng, Shubham Tulsiani, and Angela Dai. Scenefactor: Factored latent 3d diffusion for controllable 3d scene generation. *arXiv preprint arXiv:2412.01801*, 2024. 3

[9] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 3

[10] Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang, and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement. *arXiv preprint arXiv:2411.06558*, 2024. 3

[11] Zhaoxi Chen, Jiaxiang Tang, Yuhao Dong, Ziang Cao, Fangzhou Hong, Yushi Lan, Tengfei Wang, Haozhe Xie, Tong Wu, Shunsuke Saito, et al. 3dtopia-xl: Scaling high-quality 3d asset generation via primitive diffusion. *arXiv preprint arXiv:2409.12957*, 2024. 3

[12] Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024. 14

[13] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 2, 3

[14] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9): 1200–1212, 2004. 3

[15] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13142–13153, 2023. 3, 5, 15

[16] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7920, 2023. 2, 3

[17] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 3

[18] Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally. *CoRR*, 2023. 3

[19] Longtao Jiang, Zhendong Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Lei Shi, Dong Chen, and Houqiang Li. Smarteraser: Remove anything from images using masked-region guidance. *arXiv preprint arXiv:2501.08279*, 2025. 3

[20] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024. 3

[21] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 127–135, 2015. 3

[22] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4019–4028, 2021. 3

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5, 15

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5, 7

[26] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2024. 14

[27] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 17

[28] Yushi Lan, Shangchen Zhou, Zhaoyang Lyu, Fangzhou Hong, Shuai Yang, Bo Dai, Xingang Pan, and Chen Change Loy. Gaussiananything: Interactive point cloud latent diffusion for 3d generation. *arXiv preprint arXiv:2411.08033*, 2024. 3

[29] Haoran Li, Haolin Shi, Wenli Zhang, Wenjun Wu, Yong Liao, Lin Wang, Lik-hang Lee, and Peng Yuan Zhou. Dreamscene: 3d gaussian-based text-to-3d scene generation via formation pattern sampling. In *European Conference on Computer Vision*, pages 214–230. Springer, 2024. 3, 6, 7

[30] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*, 2020. 14

[31] Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3d: Fast text-to-3d with sparse-view generation and large reconstruction model. *arXiv preprint arXiv:2311.06214*, 2023. 3

[32] Weiyu Li, Jiarui Liu, Rui Chen, Yixun Liang, Xuelin Chen, Ping Tan, and Xiaoxiao Long. Craftsman: High-fidelity mesh generation with 3d native generation and interactive geometry refiner. *arXiv preprint arXiv:2405.14979*, 2024. 3

[33] Xiao-Lei Li, Haodong Li, Hao-Xiang Chen, Tai-Jiang Mu, and Shi-Min Hu. Discene: Object decoupling and interaction modeling for complex scene generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–12, 2024. 3

[34] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6517–6526, 2024. 3

[35] Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. *arXiv preprint arXiv:2402.04717*, 2024. 3

[36] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 3

[37] Huan Ling, David Acuna, Karsten Kreis, Seung Wook Kim, and Sanja Fidler. Variational amodal object completion. *Advances in Neural Information Processing Systems*, 33:16246–16257, 2020. 3

[38] Anji Liu, Mathias Niepert, and Guy Van den Broeck. Image inpainting via tractable steering of diffusion models. *arXiv preprint arXiv:2401.03349*, 2023. 3

[39] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9371–9381, 2021. 3

[40] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10072–10083, 2024. 3

[41] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023. 3

[42] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[43] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9970–9980, 2024. 3

[44] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, pages 416–423. IEEE, 2001. 7

[45] Quan Meng, Lei Li, Matthias Nießner, and Angela Dai. Lt3sd: Latent trees for 3d scene diffusion. *arXiv preprint arXiv:2409.08215*, 2024. 3

[46] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12663–12673, 2023. 3

[47] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 7, 8

[48] Başak Melis Öcal, Maxim Tatarchenko, Sezer Karaoğlu, and Theo Gevers. Sceneteller: Language-to-3d scene generation. In *European Conference on Computer Vision*, pages 362–378. Springer, 2024. 3

[49] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3931–3940. IEEE Computer Society, 2024. 3, 5, 7, 15, 16

[50] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. *Advances in Neural Information Processing Systems*, 34:12013–12026, 2021. 3

[51] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10775–10784, 2021. 3

[52] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach.

Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 7

[53] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3

[54] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 3

[55] Lu Qi, Jason Kuen, Weidong Guo, Tiancheng Shen, Jiuxiang Gu, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High-quality entity segmentation. *arXiv preprint arXiv:2211.05776*, 2022. 5, 14

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 7, 8

[57] N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9356–9366, 2022. 3

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[59] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*, 2023. 3

[60] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 3

[61] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023. 3

[62] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3

[63] Jiapeng Tang, Yinyu Nie, Lev Markhasin, Angela Dai, Justus Thies, and Matthias Nießner. Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20507–20518, 2024. 3

[64] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[65] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12619–12629, 2023. 3

[66] Yian Wang, Xiaowen Qiu, Jiageng Liu, Zhehuan Chen, Jiting Cai, Yufei Wang, Tsun-Hsuan Johnson Wang, Zhou Xian, and Chuang Gan. Architect: Generating vivid and interactive 3d scenes with hierarchical 2d inpainting. *Advances in Neural Information Processing Systems*, 37:67575–67603, 2024. 1

[67] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36:8406–8441, 2023. 3

[68] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 7, 8

[69] Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 3

[70] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2, 3, 4, 6

[71] Chao Xu, Ang Li, Linghao Chen, Yulin Liu, Ruoxi Shi, Hao Su, and Minghua Liu. Sparp: Fast 3d object reconstruction and pose estimation from sparse views. In *European Conference on Computer Vision*, pages 143–163. Springer, 2024. 3

[72] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3, 6

[73] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9099–9109, 2024. 3

[74] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 3

[75] Haitao Yang, Zaiwei Zhang, Siming Yan, Haibin Huang, Chongyang Ma, Yi Zheng, Chandrajit Bajaj, and Qixing Huang. Scene synthesis via uncertainty-driven attribute synchronization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5630–5640, 2021. 3

[76] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2024. 14

[77] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6796–6807, 2024. 3

[78] Haiyang Ying, Yixuan Yin, Jinzhi Zhang, Fan Wang, Tao Yu, Ruqi Huang, and Lu Fang. Omniseg3d: Omniversal 3d segmentation via hierarchical contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20612–20622, 2024. 2, 5, 14

[79] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024. 3

[80] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6658–6667, 2024. 3

[81] Guangyao Zhai, Evin Pınar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graph diffusion. *Advances in Neural Information Processing Systems*, 36:30026–30038, 2023. 3

[82] Guangyao Zhai, Evin Pınar Örnek, Dave Zhenyu Chen, Ruotong Liao, Yan Di, Nassir Navab, Federico Tombari, and Benjamin Busam. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, pages 167–184. Springer, 2024. 3

[83] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28003–28013, 2024. 3

[84] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3784–3792, 2020. 7

[85] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models. *ACM Transactions On Graphics (TOG)*, 42(4):1–16, 2023. 3

[86] Jingbo Zhang, Xiaoyu Li, Ziyu Wan, Can Wang, and Jing Liao. Text2nerf: Text-driven 3d scene generation with neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 30(12):7749–7762, 2024. 3

[87] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43 (4):1–20, 2024. 3

[88] Qihang Zhang, Chaoyang Wang, Aliaksandr Siarohin, Peiye Zhuang, Yinghao Xu, Ceyuan Yang, Dahua Lin, Bolei Zhou, Sergey Tulyakov, and Hsin-Ying Lee. Towards text-guided 3d scene composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6829–6838, 2024. 3

[89] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 3

[90] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in neural information processing systems*, 36:73969–73982, 2023. 3

[91] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 3

[92] Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suya You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision*, pages 324–342. Springer, 2024. 3

[93] Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. *arXiv preprint arXiv:2402.07207*, 2024. 3, 6, 7

[94] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1464–1472, 2017. 7

HiScene: Creating Hierarchical 3D Scenes with Isometric View Generation

Supplementary Material

In this supplementary material, we describe more training details of Hierarchical Scene Parsing in Sec. A and Amodal Completion in Sec. B. We provide more details on Spatial Aligned Generation in Sec. C and Sec. D. We provide detailed explanation of our user study in Sec. E. We also conducted Runtime Evaluation in Sec. F. More qualitative results can be found in our supplementary video. Source code and dataset will be released upon the acceptance of this paper.

A. Hierarchical Scene Parsing Details

A.1. Predefined Viewpoints

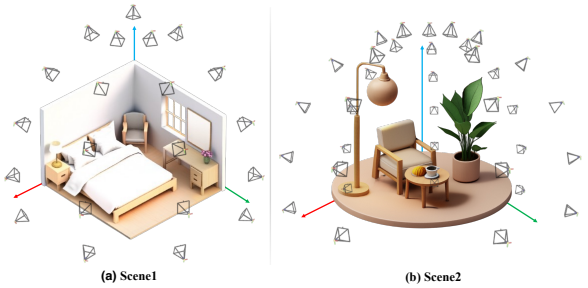


Figure 10. We provide examples of predefined viewpoints.

In the 3D Semantic Segmentation stage, we render multi-view images of the initial scene S_0 from predefined viewpoints V. As shown in the Figure 10, for scenes of type (a), we position camera viewpoints V on a sphere with a fixed radius, with azimuth angles uniformly sampled at 8 points within $[-5°, 95°]$ and elevation angles uniformly sampled at 5 points within $[-10°, 90°]$. For scenes of type (b), the azimuth angles span $[0°, 360°)$. The camera viewpoint set can be represented as $V = \{(r, \theta_i, \phi_j) \mid \theta_i \in \Theta, \phi_j \in \Phi\}$, where $r = 2$ is the sphere radius, and Θ and Φ are the sampling sets for azimuth and elevation angles, respectively. Through experimental validation, we found that the specific configuration and number of rendering viewpoints do not significantly affect 3D segmentation performance.

As shown in Figure 2 of the main paper, during the object-centric novel view synthesis stage, we fix the elevation angle at $\phi = 30°$ and uniformly select 16 viewpoints with azimuth angles $\theta \in [0°, 360°)$. All cameras are directed toward the center of the target object.

A.2. 3D Semantic Segmentation

Existing methods [12, 26, 76, 78] support 3DGS semantic segmentation, and we adopt OmniSeg3D-GS [78]. Given a collection of multi-view images $\{\mathbf{I}_i\}_{i=1}^{M}$ with their corresponding 2D semantic masks $\{\mathbf{M}_i\}_{i=1}^{M}$. OmniSeg3D-GS lifts class-agnostic 2D segmentation to 3D space through contrastive learning. To obtain complete 3D instance segmentation results, we employ EntitySeg [55] for 2D segmentation.

First, each Gaussian primitive in the initialized scene $\mathcal{S}_0$ is assigned an optimizable feature vector $\mathbf{h}_g \in \mathbb{R}^{16}$. The optimization phase follows an iterative approach, randomly selecting an image in each iteration and sampling N points from it, while determining the patch id for each point. Subsequently, differentiable rendering techniques are used to compute the render features $\{\mathbf{f_i}\}(i \in [1, N])$ for each sampled point. Within the contrastive learning framework, points sharing the same patch id are treated as positive samples, while the remaining points are considered negative samples.

To enhance computational efficiency and ensure stable convergence, OmniSeg3D-GS applies a contrastive clustering method [30]. For a set of point features sharing the same patch id i, defined as cluster $\{\mathbf{f}^i\}$, with mean feature representation $\bar{\mathbf{f}}^i$. The contrastive clustering loss is defined as:

$$\mathcal{L}_{CC} = -\frac{1}{N_p} \sum_{i=1}^{N_p} \sum_{j=1}^{|\{\mathbf{f}^i\}|} \log \frac{\exp(\mathbf{f}_j^i \cdot \bar{\mathbf{f}}^i / \phi_i)}{\sum_{k=1}^{N_p} \exp(\mathbf{f}_j^i \cdot \bar{\mathbf{f}}^k / \phi_k)}, \quad (3)$$

where $\mathbf{f}_j^i$ represents the render feature with point index j and patch id i. N_p represents the total number of patch ids, and ϕ_i is the temperature parameter for the cluster, used to balance cluster size and variance, calculated as:

$$\phi_i = \frac{\sum_{j=1}^{n_i} ||\mathbf{f}_j^i - \bar{\mathbf{f}}^i||_2}{n_i \log(n_i + \alpha)}, \quad n_i = |\{\mathbf{f}^i\}| \quad (4)$$

where $\alpha = 10$ is a smoothing parameter that prevents small clusters from producing excessively large ϕ_i values. For more details, please refer to the original paper.

A.3. Occlusion Analysis

To achieve better generation results, the image condition for LRM and TREELIS models should ideally include complete multi-view observations of the target object. In the Predefined Viewpoints setting, we render 16 candidate images around the object with a 360-degree view. We uniformly divide the 360-degree view into 4 regions $\mathcal{R} = \{R_1, R_2, R_3, R_4\}$, with each region R_i containing 4 images. For each region R_i, we randomly select one image and use a vision-language model (VLM) [1] to determine whether the object is occluded. Figure 11 shows an example prompt

used for occlusion analysis. When occlusion is detected in an image, we calculate the occlusion ratio ρ, defined as:

$$\rho = \frac{A_{\text{other}}}{A_{\text{other}} + A_{\text{target}}} \tag{5}$$

where A_{other} represents the area of other objects and A_{target} represents the area of the target object. When the occlusion ratio $\rho < 0.4$, we apply amodal completion model to process the image. Otherwise, we discard the image and randomly select another image from the same region for evaluation. If all images in a region R_i fail to meet the requirements, we proceed with only the qualifying images from other regions for subsequent processing.

B. Amodal Completion Implementation Details

B.1. Dataset Preparation

Our large-scale amodal completion dataset is derived from two sources. Part of the data comes from previous work [49], which is processed and annotated based on the SA-1B [24] dataset. Readers can refer to the original paper for more details. Additionally, to effectively handle visual effects (such as shadows) produced by occluding objects in real-world scenarios, we synthesized additional training data containing natural shadow effects.

High-quality Shadow Data Synthesis Based on Objaverse. Objaverse [15], as a large-scale diverse dataset containing over 800K 3D assets, provided us with rich 3D model resources. However, there are numerous low-quality models in this dataset, which often have overly simple geometric structures or missing textures. Therefore, we first strictly filtered Objaverse, ultimately obtaining approximately 181K high-quality 3D models for subsequent processing.

To generate realistic shadow effects, we used Blender* Cycles rendering engine for high-quality rendering. In real-world scenarios, objects are typically placed on the ground or other supporting surfaces rather than floating in space, which is crucial for shadow formation. Simple coordinate normalization cannot solve the problem of object-ground contact, so we adopted a physics simulation approach: first normalizing the 3D models to the $[-1, 1]^3$ spatial range, then adding Rigid Body Constraints to the object, and simulating the natural process of objects falling to the ground under gravity. In Figure 12, we present an example. We set the simulation time to 200 seconds and used the object's posture in the final stable state for rendering to ensure natural contact between the object and the ground.

Lighting Setup. To simulate diverse real-world lighting conditions, we primarily used two types of light sources in Blender during the rendering process: Sun Light and Area Light. Sun Light simulates parallel light rays produced by distant light sources, with controllable lighting effects through parameters such as Strength and Angle; Area Light simulates light emitted from a plane, achieving different lighting effects by adjusting parameters such as Size and Power. As outlined in Algorithm1, for each rendering, we randomly selected one of these light source types and randomly set relevant parameters to enhance data diversity.

After rendering, we adopted a method similar to previous

*https://www.blender.org

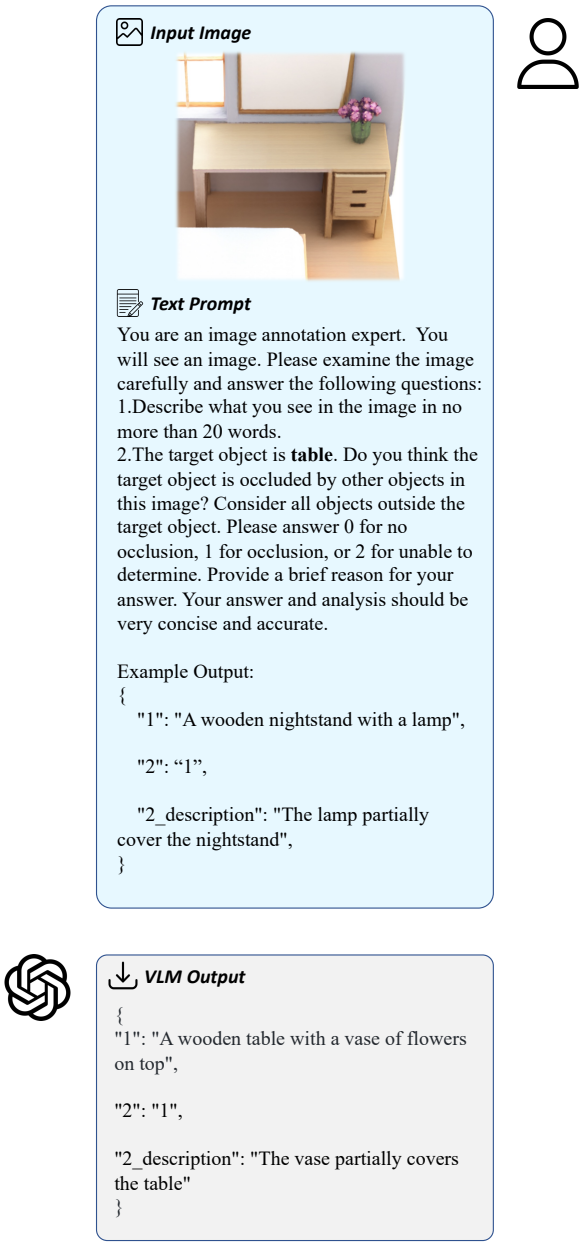
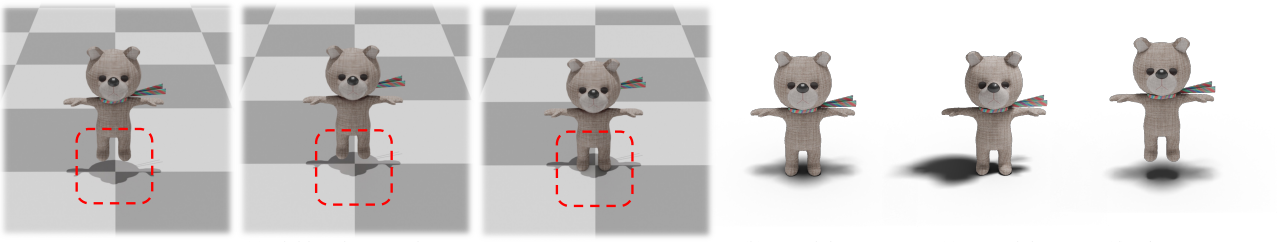


Figure 11. We provide an example of using VLM to determine whether the target object is occluded.



(a) Rigid Body Dynamics (b) Sun Light (c) Area Light (d) w/o RBD

Figure 12. (a) Visualization of the Rigid Body Dynamics (RBD) process; (b) and (c) illustrate the shadow effects under different lighting setup; (d) demonstrates that incorrect shadow effects when object are not pre-processed with RBD.

Algorithm 1 Randomized Lighting Setup

Ensure: Scene with randomized lighting configuration

Sample $p \sim \mathcal{U}(0, 1)$
if $p > 0.2$ **then**
  /* Area Light Setup */
  Sample radius $R \sim \mathcal{U}(4.0, 6.0)$
  Sample energy $E \sim \mathcal{U}(800, 1200)$
  Sample size $S \sim \mathcal{U}(0.8, 1.2)$
  Sample elevation angle $\theta \sim \mathcal{U}(40, 89.9)$
  Sample azimuth angle $\phi \sim \mathcal{U}(0, 360)$
  Add area light $\mathcal{L}(R, E, S, \theta, \phi)$
else
  /* Sun Light Setup */
  Sample sun angle $\alpha \sim \mathcal{U}(0.1, 0.5)$
  Add primary sun light with angle α and energy 5.0
  for $i \in \{1, 2, 3\}$ **do**
    Add sun light with angle α and energy 3.0
    Rotate light by $\{90, 180, 270\}[i-1]$ around x-axis
  end for
end if

Algorithm 2 Occlusion-to-Visibility Transition Generation

Require: M_v: visible mask, M_w: whole mask, I_o: target RGB image, N: interpolation frames
Ensure: $\mathbf{X}$: video data

$M_d \leftarrow |M_v - M_w|$ {Difference mask}
$I_v \leftarrow I_o \odot M_w$ {Visible RGB}
Initialize $\mathbf{X} \in \mathbb{R}^{(N+1)\times c\times h\times w}$ {Output tensor}
$\mathbf{X}_0 \leftarrow \text{normalize}(\text{resize}(I_o, w \times h))$ {First frame}
for $i \in \{0, 1, ..., N-1\}$ **do**
  $\alpha_i = 1 - \frac{i}{N-1}$ {Blending coefficient}
  $M_i \leftarrow M_d \cdot \alpha_i$ {Weighted mask}
  $I_i \leftarrow \text{blend}(I_v, I_w, M_i)$ {Blended image}
  $\mathbf{X}_{i+1} \leftarrow \text{normalize}(\text{resize}(I_i, w \times h))$
end for
return $\mathbf{X}$

method [49], overlaying 3D object images with shadows onto complete scene images from SA-1B to form image pairs containing occlusion relationships. Through the above processing, we ultimately constructed a large-scale amodal completion dataset containing approximately 1.32 million data pairs, of which about 468K data pairs include natural shadow effects, providing rich training resources for the model to learn to process complex occlusion scenarios. As shown in Figure 12, the shadow effects under different lighting are illustrated.

B.2. Network Architecture

In Section 4, we proposed a amodal completion network based on Stable Video Diffusion [7] (SVD) . This section elaborates on its network architecture design.

Stable Video Diffusionis an image-conditioned video generation network. Trained on massive video datasets, given an input image I, it can generate a temporally consistent video from I. The SVD architecture consists of three core components: a video encoder $\mathcal{E}$, a decoder $\mathcal{D}$, and a UNet diffusion model Φ.

To fully leverage the prior knowledge embedded in the pre-trained model, we optimize the network structure following the principle of minimal modification. Specifically, the input image I is first compressed through the encoder $\mathcal{E}$ to obtain an image embedding representation. Simultaneously, the object observation mask M undergoes downsampling to maintain consistent resolution with the image embedding. In the original SVD architecture, the image embedding is concatenated with latent noise; in our improved version, we incorporate the mask M into the concatenation process, providing additional information to better guide the video generation process. The image I is also processed through CLIP to extract image features, which are injected via cross-attention.

B.3. Training Details

In the model training process, we initialized our model with the pre-trained weights from SVD [7]. To accommodate the additional mask information, we expanded the input

Algorithm 3 Spatial Aligned Generation

Require: Observation views I with camera parameters $[r_i, p_i]$, timestep range $[t_{min}, t_{max}]$
Ensure: Generated 3D structure $\widehat{S}$

/* Initial Structure Generation */
Generate initial geometric structure $S_{\text{init}} \leftarrow F_{\text{LRM}}(\text{I}, [r_i, p_i])$
Obtain explicit voxel representation V through voxelization of S_{init}
Compress V into latent feature $S_{\text{implicit}} \leftarrow \mathcal{E}_S(V)$
/* Shape Prior Injection */
Select intermediate timestep $t \in [t_{min}, t_{max}]$ {Typically $t \in [0.2, 0.4]$}
Initialize $x_t \leftarrow S_{\text{implicit}} + \text{noise}(t)$ {Add noise corresponding to timestep t}
Set number of discretization steps N
/* Rectified Flow Sampling Process */
for $i = \text{INT}(N \times t - 1)$ **down to** 0 **do**
    Compute current time t_i, step size h
    Predict velocity $v_\theta(x_{t_i}, t_i)$ using trained model $\boldsymbol{\mathcal{G}}_{\text{S}}$
    Update sample $x_{i-1} = x_i - h \cdot v_\theta(x_i, t_i)$
end for
Set generated structure $\widehat{S} \leftarrow x_0$
return $\widehat{S}$

channels of thn illustration of the use diffusion model Φ from 8 to 9. In our experiments, video data was processed at a resolution of 512×512 with 8 frames per sequence. Training was conducted using fp16 precision with a learning rate of 1×10^{-5} and incorporated a 500-step warmup phase. All parameters of Φ were fine-tuned during this process. We used a batch size of 8 and trained on 8 Nvidia H20 GPUs, with the entire training process taking approximately 3 days.

C. Spatial Aligned Generation

Here we provide an algorithm 3 for spatial aligned generation.

D. Background Structure Generation

For each scene, we regenerate the background by modifying the original scene prompt to begin with 'A background of...', then integrate the resulting background structure back into the resulting scene.

E. User Study

Our user study evaluates the generated 3D scenes from two dimensions: image-text matching degree and overall scene quality. The matching degree assesses how well the generated 3D scene aligns with the input text description, while the overall quality comprehensively considers factors such as scene quality, reasonableness of object layout, and the quality of individual object models. As shown in Figure 13, we designed an intuitive user study interface where participants can quantitatively score the generated results. The scoring standard ranges from 1 to 3 points, with higher scores indicating better quality. Through this approach, we can qualitatively analyze the practical effectiveness of the generated results.

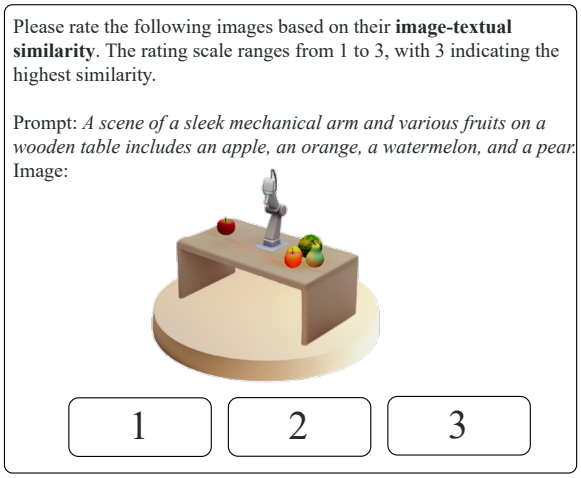


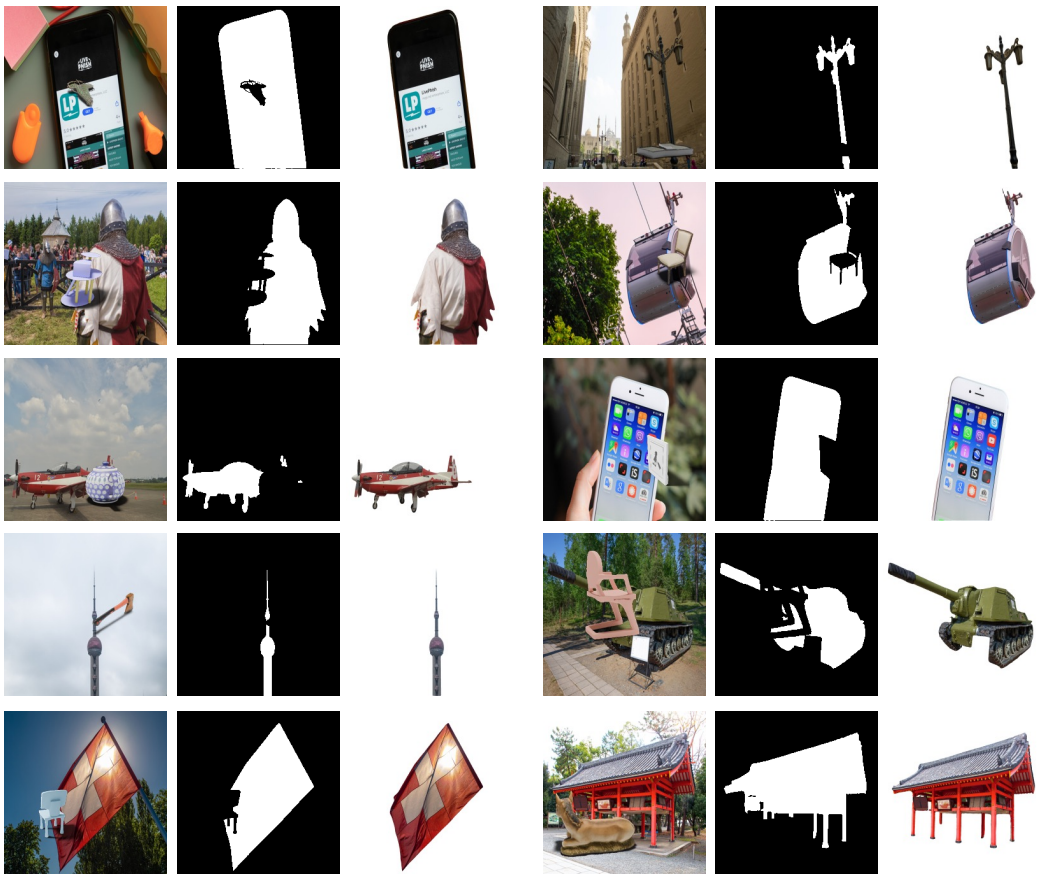
Figure 13. An illustration of the user study interface.

F. Runtime Evalution

As shown in Figure 2 of the main paper, given a text prompt, HiScene first utilizes FLUX [27] to generate candidate images within 4 seconds, followed by TRELLIS to complete scene initialization in 5 seconds. Next, HiScene renders multi-view images and performs 2D semantic segmentation in 5 seconds. In the 3D Gaussian semantic segmentation stage, we optimize 5000 steps, taking approximately 2 minutes. After obtaining the initial 3D GS representation of objects, the system spends less than 1 second rendering object candidate images and completes occlusion analysis in 30 seconds. For each candidate image with occlusion, amodal completion requires 6 seconds. As illustrated in Figure 5 of the main paper, during the Spatial Alignment stage for each object, we first utilize Sparse view LRM to obtain initial geometric structures in 2 seconds, followed by voxelization in less than 1 second, and generate a low-resolution feature grid using E_S in 1 second. Finally, the system takes 3 seconds for structure generation and structure latents generation.

Overall, HiScene processes a complete scene in approximately 12 minutes. In contrast, methods relying on SDS Loss optimization such as GALA3D and DreamScene require significantly more time. GALA3D needs 2 hours to generate a scene, while DreamScene requires more than 1 hour. This comparison clearly demonstrates HiScene's significant advantage in efficiently generating interactive 3D scenes.

(a) Rendered shadow-aware Object data



(a) Synthetic shadow-aware image pairs

Figure 14. We show examples of our synthetic dataset.

"A cozy, elegant scene reveals a circular platform featuring a classic upright piano, a green upholstered armchair, a matching red piano stool, and a tall standing lamp."

"A minimalist workspace. A light wooden desk with three drawers holds an open laptop and a black desk lamp. A black swivel chair sits on a blue mat, and a gray trash bin is placed nearby"

"A minimalist workspace. A light wooden desk with three drawers holds an open laptop and a black desk lamp. A black swivel chair sits on a blue mat, and a gray trash bin is placed nearby."

"A modern dining area reveals a rectangular marble-patterned table, four white chairs, a pitcher, small bowls, and a wooden circular platform."

"A dining room features a long wooden table surrounded by chairs."

Figure 15. More examples of generated scenes. All prompts have a fixed prefix *"Isometric view of "*.

Figure 16. More examples of amodal completion.